There are good reasons to randomize the assignment of $\vec{w}$
$\Rightarrow$ the "randomized experiment" (1925, Fisher)
$\Rightarrow \vec{w} \sim \frac{1}{\binom{n}{n_T}} \mathbb{1}_{\sum w_i = n_T}$ is called the "completely randomized design".

$E[\vec{w}] = \frac{1}{2}\vec{1}_n$. This design has the added bonus that the bias disappears:

$$\beta_T = \frac{2\vec{w}^T\vec{Y} - \vec{1}^T\vec{Y}}{\frac{n}{2}} = \frac{4}{n}\vec{w}^T(\beta_0\vec{1}_n + \beta_u\vec{u} + \beta_T\vec{w} + \vec{\varepsilon}) - \frac{2}{n}\vec{1}^T(\beta_0\vec{1}_n + \beta_u\vec{u} + \beta_T\vec{w} + \vec{\varepsilon})$$

$$= \frac{4}{n}\left(\beta_0 \underbrace{\vec{1}^T\vec{w}}_{\frac{n}{2}} + \beta_u\vec{u}^T\vec{w} + \beta_T\underbrace{\vec{w}^T\vec{w}}_{\frac{n}{2}} + \vec{w}^T\vec{\varepsilon}\right) - \frac{2}{n}\left(\beta_0\underbrace{\vec{1}_n^T\vec{1}_n}_{n} + \beta_u\vec{1}_n^T\vec{u} + \beta_T\underbrace{\vec{1}^T\vec{w}}_{\frac{n}{2}} + \vec{1}^T\vec{\varepsilon}\right)$$

$$= 2\beta_0 + \frac{4\beta_u}{n}\vec{u}^T\vec{w} + 2\beta_T + \frac{4}{n}\vec{w}^T\vec{\varepsilon} - 2\beta_0 - \frac{2\beta_u}{n}\vec{1}_n^T\vec{u} - \beta_T - \frac{2}{n}\vec{1}^T\vec{\varepsilon}$$

$$E_{\vec{\varepsilon}}[\,\cdot\,] = \beta_T + \frac{4\beta_u}{n}\vec{u}^T\vec{w} - \frac{2\beta_u}{n}\vec{1}_n^T\vec{u}$$

$$E_{\vec{w}}[\,\cdot\,] = \beta_T + \frac{2}{\cancel{4}}\frac{4\beta_u}{n}\vec{u}^T\left(\frac{1}{2}\vec{1}_n\right) - \frac{2\beta_u}{n}\vec{1}_n^T\vec{u} = \beta_T \checkmark$$

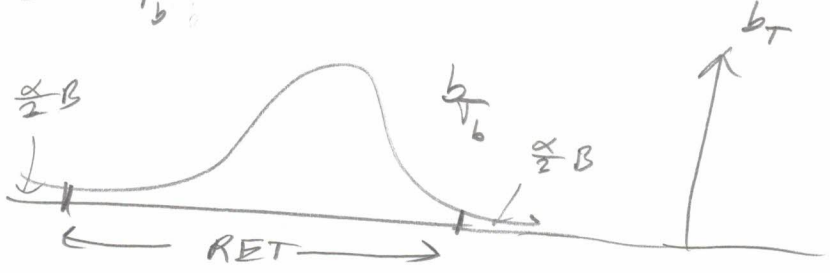
On average, over all randomized experiments, the PATE is measured unbiasedly

There is a whole other reason to employ randomization. Fisher called it a "reasoned basis for inference". Here's how this works.

Let H_0: $Y_i(w_i=1) = Y_i(w_i=0)$ $\forall i$ that is, the response value will be identical regardless of the manipulation

You run the the experiment with $\vec{w}_{exp}$ and collect $\vec{y}$ and compute b_T.

Then for B = large #, you generate $\vec{w}$ from $\vec{w}$ and compute b_{T_b}



Then you can form null limits at the $\frac{\alpha}{2}$ and $1-\frac{\alpha}{2}$ empirical quantiles.

If b_T from the actual experiment falls outside of RET $\Rightarrow$ reject H_0.

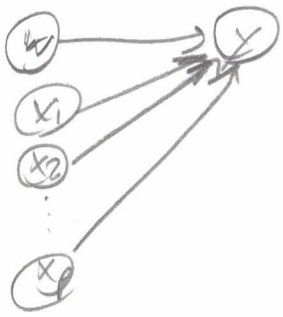
Note: no assumptions about the population model such as $\vec{\varepsilon} \sim N_n(\vec{0}, \sigma^2 I_n)$ are necessary! This is a non-parametric test just like the perm. test.

$\Rightarrow$ Randomization gives you:

(1) insurance against unmeasured covariate or experimental impurity

(2) the ability to have a non-parametric test option

What about the classic baseline covariates?



let $\vec{Y} = \beta_0 \vec{1}_n + \beta_T \vec{w} + \beta_1 \vec{x}_1 + \ldots + \beta_p \vec{x}_p + \vec{\varepsilon}$

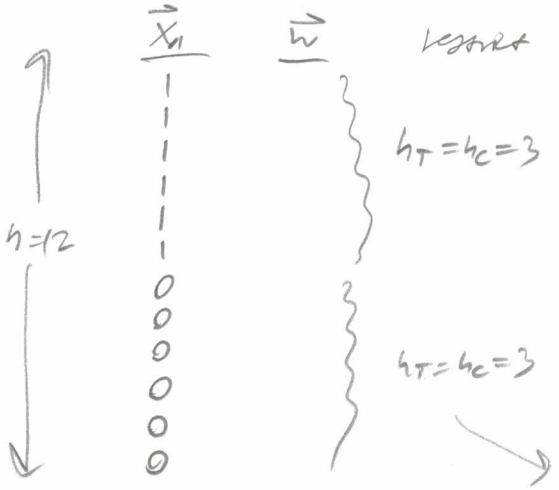
Similar to comparison from before...

$$\Rightarrow \hat{\beta}_T = \bar{Y}_T - \bar{Y}_C = \beta_T + \beta_1(\bar{x}_{1T} - \bar{x}_{1C}) + \beta_2(\bar{x}_{2,T} - \bar{x}_{2C}) + \ldots + \beta_p(\bar{x}_{p,T} - \bar{x}_{p,C}) + \bar{\varepsilon}_T - \bar{\varepsilon}_C$$

$$\Rightarrow E_{\vec{\varepsilon}}[\hat{\beta}_T] = \beta_T + \underbrace{\beta_1(\bar{x}_{1,T} - \bar{x}_{1,C}) + \ldots + \beta_p(\bar{x}_{p,T} - \bar{x}_{p,C})}_{\text{Bias}}$$

We can reduce this bias by "restricting" the randomization to keep these terms small. This was noted by Fisher immediately.

For example, let x_1 = gender and let $n = 12$



$$\Rightarrow \bar{x}_{1,T} = 0.5, \quad \bar{x}_{1,C} = 0.5$$

$$\Rightarrow \beta_1(\bar{x}_{1,T} - \bar{x}_{1C}) = 0 \text{ for all } \vec{w}!!$$

→ Only choose $\vec{w}$'s that respect this constraint!

This is known as "blocking". The two blocks above are one for male, one for female.

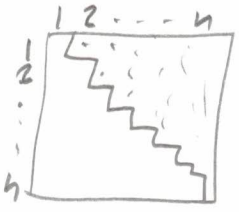
If x_1 is continuous ⇒ balance or cut in 3 or 4 even pieces (blocks)
then randomize with equal allocation within block

Blocking is the most common restricted design.

However blocking doesn't work for p large as block size is $\frac{n}{2^p} \to 0$ quickly. My research now shows that a good recommendation is the pairwise matching design: split the n subjects into $\frac{n}{2}$ pairs.

How? Define Euclidean distance: $d(\vec{x}_{ki}, \vec{x}_{\ell i}) = \sum_{j=1}^{p} (x_{kj} - x_{\ell j})^2$

Create distance matrix:



upper triangular $D = \ell$

Then use the nonbipartite matching algorithm to create pairs that minimize the overall distance of total $\sum_{k,\ell} d_{k,\ell}$. Within pairs, randomize $w \in \{0,1\}$

You can think of this as $\frac{n}{2}$ blocks of size 2 each.

Another design that has gotten attention: rerandomization (Student's idea). Draw a $\vec{w}$. If $\sum_{j=1}^{p} (\bar{x}_{j,T} - \bar{x}_{j,C})^2$ is "large", draw another $\vec{w}$ until it's small i.e. $\leq D_{th}$, your hyperparameter.

For blocking, pairwise matching, rerandomization, you can use the randomization test as well... just make sure you draw $\vec{w}$'s that are legal within the restricted allocation set.

What if $y = \{C_1, C_2, \ldots, C_K\}$ classification with K levels.

If $K=2$, let's review... assume $Y_i \overset{ind}{\sim} Bern(\phi(\vec{x}_i\vec{\beta}))$ where ϕ is a link function $\phi: \mathbb{R} \to (0,1)$.

$\Rightarrow$ Lik. Function:

$$\mathcal{L}(\vec{\beta}; X, \vec{y}) = \prod_{i=1} \phi(\vec{x}_i\vec{\beta})^{y_i} (1-\phi(\vec{x}_i\vec{\beta}))^{1-y_i}$$

if $\phi(\eta) = \frac{e^\eta}{1+e^\eta}$, logistic regression

If $K>2$, assume $Y_i \overset{ind}{\sim} Multi(1, \vec{\theta}_i) = \binom{1}{y_{i1} y_{i2} \cdots y_{iK}} \theta_1^{y_{i1}} \theta_2^{y_{i2}} \cdots \theta_K^{y_{iK}}$

let $\vec{\theta}_i = \begin{bmatrix} \phi(\vec{x}_i\vec{\beta}_1) \\ \phi(\vec{x}_i\vec{\beta}_2) \\ \vdots \\ \phi(\vec{x}_i\vec{\beta}_K) \end{bmatrix}$ and since $\vec{1}_K^T \vec{\theta} = 1$, $\phi(\vec{x}_i\vec{\beta}_K) = 1 - (\phi(\vec{x}_i\vec{\beta}_1) + \cdots + \phi(\vec{x}_i\vec{\beta}_{K-1}))$

$\Rightarrow$ Likelihood Function is:

$$\mathcal{L}(\vec{\beta}_1, \vec{\beta}_2, \ldots, \vec{\beta}_{K-1}; X, \vec{y}) = \prod_{i=1} \phi(\vec{x}_i\vec{\beta}_1)^{y_{i,1}} \phi(\vec{x}_i\vec{\beta}_2)^{y_{i,2}} \cdots \phi(\vec{x}_i\vec{\beta}_{K-1})^{y_{i,K-1}} \cdot \left(1 - (\phi(\vec{x}_i\vec{\beta}_1) + \cdots + \phi(\vec{x}_i\vec{\beta}_{K-1}))\right)^{y_{i,K}}$$

If $\phi(\eta) = \frac{e^\eta}{1+e^\eta} \Rightarrow$ multilogit model

If $\phi(\eta) = \Phi(\eta) \Rightarrow$ multiprobit model

} most common

Inference? Use Wald test. Who is the effect of x_1? There are $K-1$ effects:

$b_{1,1}, b_{2,1}, \ldots b_{K-1,1}$ where each is a log-odds effect on prob of class k!

Anyway to interpret... but you get used to it.